

## DOCUMENT RESUME

ED 351 383

TM 019 215

AUTHOR Nandakumar, Ratna; Stout, William  
 TITLE Refinements of Stout's Procedure for Assessing Latent Trait Unidimensionality.  
 INSTITUTION Illinois Univ., Urbana. Dept. of Statistics.  
 SPONS AGENCY Office of Naval Research, Arlington, Va.  
 REPORT NO 1992-1; ONR-4421-548  
 PUB DATE 1 Aug 92  
 CONTRACT NO00014-90-J-1940  
 NOTE 52p.; Paper to be published in the "Journal of Educational Statistics."  
 PUB TYPE Reports - Evaluative/Feasibility (142)  
 EDRS PRICE MF01/PC03 Plus Postage.  
 DESCRIPTORS Computer Simulation; \*Computer Software; Equations (Mathematics); Estimation (Mathematics); Guessing (Tests); Hypothesis Testing; \*Item Response Theory; \*Mathematical Models; \*Psychological Testing; Statistical Significance; \*Test Items  
 IDENTIFIERS American College Testing Program; Armed Services Vocational Aptitude Battery; Binary Response; DIMTEST (Computer Program); \*Stouts Procedure; \*Unidimensionality (Tests)

## ABSTRACT

A detailed investigation of the statistical procedure of W. Stout (the computer program DIMTEST) for testing the hypothesis that an essentially unidimensional latent trait model fits observed binary response data from a psychological test is presented. One finding is that DIMTEST may fail to perform as desired in the presence of guessing when coupled with many high-discriminating test items. A revision of DIMTEST is proposed to overcome this limitation. Also, an automatic approach is devised to determine the size of the assessment subtests. Further, an adjustment is made on the estimated standard error of the statistic on which DIMTEST depends. These three refinements result in an improved procedure that is shown in simulation studies to adhere closely to the nominal level of significance while achieving considerably greater power. Finally, DIMTEST is validated on real data sets from the Armed Services Vocational Aptitude Battery (1,984 and 1,961 examinees) and the American College Testing program (2,491 and 2,494 examinees). Seven tables present analysis results, and 46 references are included.  
 (Author/SLD)

\*\*\*\*\*  
 \* Reproductions supplied by EDRS are the best that can be made \*  
 \* from the original document. \*  
 \*\*\*\*\*

83  
83  
51  
33  
D  
E

## Refinements of Stout's Procedure for Assessing Latent Trait Unidimensionality

Ratna Nandakumar<sup>1</sup>

Department of Educational Studies  
University of Delaware

William Stout

Department of Statistics  
University of Illinois

August 1, 1992

Prepared for the Cognitive Science Research Program, Cognitive and Neural Sciences Division, Office of Naval Research, under grant number N00014-90-J-1940, 4421-548. Approved for public release, distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

---

<sup>1</sup> The authors would like to convey special thanks for Brian Junker for his time and insightful comments on earlier drafts of this manuscript; to Mark Reckase and Robert Linn for providing us with the real data sets; and to Roderick McDonald for allowing us to use the nonlinear factor analysis program NOFA.

# REPORT DOCUMENTATION PAGE

*Form Approved  
OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED	
	1 August 1992	Technical: 1990-93	
4. TITLE AND SUBTITLE <b>Refinements of Stout's Procedure for Assessing Latent Trait Unidimensionality</b>		5. FUNDING NUMBERS <b>N00014-90-J-1940</b>	
6. AUTHOR(S) <b>Ratna Nandakumar and William Stout</b>			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) <b>Department of Statistics University of Illinois 725 South Wright Street Champaign, IL 61820</b>		8. PERFORMING ORGANIZATION REPORT NUMBER <b>1992 - No. 1</b>	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) <b>Cognitive Sciences Program Office of Naval Research 800 North Quincy Arlington, VA 22217-5000</b>		10. SPONSORING / MONITORING AGENCY REPORT NUMBER <b>NR 4421-548</b>	
11. SUPPLEMENTARY NOTES <b>To be published in Journal of Educational Statistics. Software to carry out procedures available from authors.</b>			
12a. DISTRIBUTION / AVAILABILITY STATEMENT <b>Approved for public release; distribution unlimited</b>		12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words) <b>See reverse</b>			
14. SUBJECT TERMS <b>See reverse</b>		15. NUMBER OF PAGES <b>47</b>	
		16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT <b>unclassified</b>	18. SECURITY CLASSIFICATION OF THIS PAGE <b>unclassified</b>	19. SECURITY CLASSIFICATION OF ABSTRACT <b>unclassified</b>	20. LIMITATION OF ABSTRACT

**Refinements of Stout's Procedure for Assessing  
Latent Trait Unidimensionality**

**Abstract**

This paper provides a detailed investigation of Stout's statistical procedure (the computer program DIMTEST) for testing the hypothesis that an essentially unidimensional latent trait model fits observed binary item response data from a psychological test. One finding was that DIMTEST may fail to perform as desired in the presence of guessing when coupled with many high-discriminating items. A revision of DIMTEST is proposed to overcome this limitation. Also, an automatic approach is devised to determine the size of the assessment subtests. Further, an adjustment is made on the estimated standard error of the statistic on which DIMTEST depends. These three refinements have led to an improved procedure that is shown in simulation studies to adhere closely to the nominal level of significance while achieving considerably greater power. Finally, DIMTEST is validated on a selection of real data sets.

Subject terms: Unidimensionality, essential independence, essential unidimensionality, DIMTEST, item response theory.

**Refinements of Stout's Procedure for Assessing  
Latent Trait Unidimensionality**

Item response theory (IRT) is presently one of the most widely used techniques in psychometrics and is likely to remain so in the future. Some applications of IRT include ability estimation, item/test bias, equating, and adaptive testing. The three assumptions underlying many commonly used IRT models are monotonicity, unidimensionality ( $d=1$ ), and local independence (LI). Monotonicity assumes that the probability of correctly answering an item increases as ability increases. Unidimensionality<sup>1</sup> assumes that the items of a test measure a single ability. Local independence assumes that given any particular level of ability, responses to different items are independent. This paper is concerned with the statistical assessment of the assumption of unidimensionality. Most IRT models specifically require this assumption; moreover, classical test theory models implicitly assume that items measure the same dominant dimension. In spite of the importance of this assumption, it is also well known that actual data are rarely strictly unidimensional. It has long been argued that items are multiply determined and that, in addition to measuring the intended attribute, other attributes unique to individual items or common to relatively few items are unavoidable (Humphreys, 1981, 1985, 1986; Hambleton & Swaminathan, 1985; Reckase, 1979, 1985; Stout, 1987; Traub, 1983; Yen, 1985). In addition to the multiple item attributes that influence dimensionality, examinee characteristics such as differential teaching methods, the point of time during the instructional unit that the test is given, and so forth, also can influence the dimensionality of a set of items (Birenbaum & Tatsuoka, 1982; Bejar, 1983; Traub, 1983). Dimensionality is therefore a property of both the test and the examinee population taking the test (Reckase, 1990).

Linear factor analysis (subjectively interpreted in the absence of a statistical distribution theory) has been the traditional approach for assessing the dimensionality of a

set of items. If the results of a linear factor analysis reveal only one significant factor, then the test is considered unidimensional. In the case of dichotomous data, however, it is well known that linear factor analysis of phi correlations between items often leads to overestimation of the number of factors underlying the item responses (Carroll, 1945; Hambleton & Swaminathan, 1985, Chapter 2; Hulin, Drasgow, & Parsons, 1983, Chapter 8; McDonald & Ahlswed, 1974). As a corrective alternative, tetrachoric correlations can be used for factor analysis. When guessing is present in the responses to items, however, linear factor analysis of tetrachoric correlations can produce a spurious factor due to difficulty of test items (Hulin, Drasgow, & Parsons, 1983, Chapter 8). In addition, computation of tetrachoric correlations can be problematic if any one of the correct/incorrect cells of the two-by-two item response tables contains a zero. Matrices of simple tetrachoric correlations are thus often non-gramian. As a result, conventional methods of factor analysis by phi or tetrachoric correlations are often unsatisfactory for assessing the dimensionality of test items. Christoffersson (1975) and Muthen (1978) have developed generalized least squares methods to overcome the problems with factor analysis of tetrachoric correlations, but their methods are limited to 25 items at most. Moreover, they are computationally intensive.

In recent years a vast body of literature has been developed for assessing the dimensionality of test items. Comprehensive reviews of different procedures for assessing dimensionality are provided by Hattie (1984, 1985), and Hulin, Drasgow, and Parsons (1983, Chapter 8). Some of the more recent procedures developed to assess latent trait dimensionality include: maximum likelihood full information factor analysis (Bock, Gibbons, & Muraki, 1985); the Tucker and Humphreys procedures based on local independence and first and second factor loadings (Roznowski, Tucker, & Humphreys, 1991); Stout's (1987) procedure for assessing unidimensionality based on the theory of essential independence; modified parallel analysis, which combines latent trait methods and

factor analysis and uses eigenvalues of tetrachoric correlation matrix (Hulin, Drasgow, & Parsons, 1983, p. 255); McDonald's nonlinear factor analysis (McDonald, 1962; McDonald & Ahlawat, 1974; Etazadi-Amoli & McDonald, 1983); Holland and Rosenbaum's test of unidimensionality, monotonicity, and conditional independence (Holland, 1981; Holland & Rosenbaum, 1986); residual analysis, determined by model-data fit (Hambleton & Swaminathan, 1985, Chapter, 8); and Bejar's procedure based on three-parameter logistic item parameter estimates (Bejar, 1980). Some of these methods are reviewed in Hambleton and Rovinelli (1986), Mislevy (1986), and Zwick (1987a).

Although these different approaches offer promise for assessing the dimensionality of binary data, researchers in the field have not reached a consensus on one satisfactory method (Berger & Knol, 1990; Hambleton & Rovinelli, 1986; Hattie, 1984; Zwick, 1987b). Primarily, this is due to the fact that there is substantial confusion in the literature concerning the definition of unidimensionality. Additionally, many existing methods for assessing dimensionality are only loosely connected to the various definitions in the literature (Hambleton & Rovinelli, 1986).

This article is concerned with Stout's procedure for assessing unidimensionality (DIMTEST). Stout (1987) has developed a nonparametric statistical procedure based on the large sample distribution theory for assessing latent trait dimensionality and has argued the validity of this procedure based upon simulation studies involving a variety of achievement tests. DIMTEST has been shown to discriminate well between one- and two-dimensional tests, maintaining good adherence to a specified level of significance when  $d=1$  and maintaining good power when  $d=2$ , even when the correlation between the abilities is as high as .7.

The present study provides a detailed investigation of certain performance characteristics and the consequent major refinements of DIMTEST for assessing latent trait dimensionality. DIMTEST was found to perform undesirably in certain cases where

the test contained many highly discriminating items with guessing present. A correction is proposed to overcome this limitation. In addition, an automatic approach is devised for determining  $M$ , the size of the assessment subtests; a better control of  $\alpha$ , the specified level of significance, is achieved by adjusting the estimated standard error of Stout's statistic  $T$ . These refinements have led to an improved test procedure that is easier to use and has been shown in simulation studies to adhere closely to the nominal level of significance while achieving considerably greater power. Finally, the procedure is applied to a selection of real data sets.

### Stout's Procedure for Assessing Unidimensionality

As stated in the beginning of this paper, items are multiply determined, and thus the number of *dominant* abilities should be assessed in testing for dimensionality. Stout first informally (1987) and then formally (1990) provided a definition of the number of dominant dimensions known as *essential dimensionality*, which is derived from the theory of *essential independence*. The statistical procedure for assessing *essential unidimensionality* is consistent with the definition of essential dimensionality. To assist the reader in evaluating this claim as well as to enable the reader in understanding the refinements made to DIMTEST, Stout's definition of essential dimensionality will be followed by a brief summary of the statistical procedure. The reader is advised, however, that use of DIMTEST does not require acceptance of Stout's notion of essential dimensionality, and, in fact, DIMTEST can also be viewed as a technique to detect sizable lack of fit of a locally independent unidimensional latent trait model.

Let  $U_i$  denote the  $i$ -th item response and  $\underline{U}_N \equiv (U_1, U_2, \dots, U_N)$ , denote the test response vector for an  $N$ -item test. Observed item and test values will be denoted by  $u_i$  and  $\underline{u}_N \equiv (u_1, u_2, \dots, u_N)$ , respectively. Let  $U_i = 1$  denote a correct response and  $U_i = 0$

denote an incorrect response to item  $i$  for a randomly chosen examinee. The latent random vector is denoted by  $\underline{\Theta}$  and the particular values it takes are denoted by  $\underline{\theta}$ . Let  $P_i(\underline{\theta})$  denote the probability that a randomly chosen examinee with ability  $\underline{\theta}$  will get the  $i$ -th item correct. It is assumed that all item response functions  $P_i(\underline{\theta})$  are monotone. Let  $\underline{U} = \{U_i, i \geq 1\}$  denote the item pool consisting of  $U_N$  as its first  $N$  items. The item pool is conceptualized as a result of continuing the test construction process in the same manner beyond the construction of the  $N$  items that make up the actual test  $U_N$  being studied. One advantage of using only the partially observed  $\underline{U}$  instead of the actually observed  $U_N$  to model the test is that a totally rigorous definition of the number of dominant dimensions can be given. These ideas are carefully and formally developed in Stout (1990) and constitute a large sample approach to test modeling.

**Definition 1** (Stout, 1990) The item pool  $\underline{U}$  is said to be *essentially independent* (EI) with respect to the latent variable  $\underline{\Theta}$  if  $\underline{U}$  satisfies

$$D_N(\underline{\theta}) \equiv \frac{\sum_{1 \leq i < j \leq N} |\text{Cov}(U_i, U_j | \underline{\Theta} = \underline{\theta})|}{\binom{N}{2}} \rightarrow 0 \text{ as } N \rightarrow \infty \quad (1)$$

for every  $\underline{\theta}$ .

The distinction between local independence and essential independence is that local independence requires  $\text{Cov}(U_i, U_j | \underline{\Theta} = \underline{\theta}) = 0$  for all  $\underline{\theta}$ ; whereas, essential independence requires the average value of  $|\text{Cov}(U_i, U_j | \underline{\Theta} = \underline{\theta})|$  over all item pairs to be small in magnitude for all  $\underline{\theta}$  as the test length increases. Hence, essential independence is a weaker assumption than local independence.

**Definition 2** (Stout, 1990). The *essential dimensionality* ( $d_E$ ) of an item pool  $\mathcal{U}$  is the minimal dimensionality (number of elements in  $\theta$ ) necessary to satisfy the assumption of essential independence. When  $d_E = 1$ , *essential unidimensionality* is said to hold.

The reader should note that  $d_E=1$  means that  $\mathcal{U}$  has an IRT model for which essential independence holds for a unidimensional latent trait  $\theta$ . The ordinary definition of IRT dimensionality is the same as Definition 2, but essential independence is replaced by local independence and  $\mathcal{U}$  replaced by  $\mathcal{U}_N$ . Stout argues that the assumptions concerning local independence and the resulting ordinary IRT definition of dimensionality should often be replaced by the respective weaker assumptions concerning essential independence and essential dimensionality. Junker (1988, 1991) has proved results concerning essential independence and, in particular, has derived statistical consistency results for maximum likelihood estimates of ability under the assumption of essential unidimensionality.

It can now be clearly stated what assessing the hypothesis of essential unidimensionality means: among all the essentially independent monotone IRT models for  $\mathcal{U}$ , does there exist a unidimensional one? To answer this question, we assume both monotonicity and essential independence and assess the lack of fit of unidimensionality. This approach is similar to most other procedures for assessing data dimensionality, with the exception that essential rather than local independence is assumed.

The statistical procedure for testing the null hypothesis of essential unidimensionality will be briefly described here. For further details see Stout (1987, Sec. 4). The  $N$  test items are split into two assessment subtests of length  $M$  each—called the Assessment 1 subtest (AT1) and the Assessment 2 subtest (AT2)—and a longer subtest called the partitioning subtest (PT) of length  $n (= N-2M)$ . The  $M$  items for subtest AT1 are selected to have the same dominant trait. This splitting can be done using either expert opinion or exploratory factor analysis. Whatever method used to select items of AT1, the

goal is to select a small subset of items (up to one-fourth of the total test length seems a good convention) that all measure the same dominant trait and, at the same time, are as dimensionally different as possible from the PT items. Once items for AT1 are selected, a second set of  $M$  items is selected for AT2 from the remaining items so that AT2 items have a difficulty distribution similar to AT1 items (Step 6, Stout, 1987). The remaining  $n$  ( $= N-2M$ ) items then become the partitioning subtest PT.

Each examinee is assigned to one of  $K$  subgroups according to his/her score on PT. After eliminating subgroups with too few examinees ( $J_{\min} = 20$  recommended), within each subgroup,  $k$ , two variance estimates, the usual variance estimate ( $\hat{\sigma}_k^2$ ), and the "unidimensional" variance estimate ( $\hat{\sigma}_{U,k}^2$ ), are computed using items of AT1.

$$\hat{\sigma}_k^2 = \sum_{j=1}^{J_k} (\bar{Y}_j^{(k)} - \bar{Y}^{(k)})^2 / J_k$$

where

$$\bar{Y}_j^{(k)} = \sum_{i=1}^M U_{ijk} / M \text{ and } \bar{Y}^{(k)} = \sum_{j=1}^{J_k} \bar{Y}_j^{(k)} / J_k$$

with  $U_{ijk}$  denoting the response of the  $j$ th examinee to the  $i$ th item from the  $k$ th subgroup, and  $J_k$  denoting the number of examinees in the  $k$ th subgroup.

$$\hat{\sigma}_{U,k}^2 = \sum_{i=1}^M \hat{p}_i^{(k)} (1 - \hat{p}_i^{(k)}) / M^2,$$

where

$$\hat{p}_i^{(k)} = \sum_{j=1}^{J_k} U_{ijk} / J_k$$

The difference in these variance estimates is then normalized by an appropriate normalizing constant  $S_k$  and summed over subgroups to arrive at the statistic

$$T_L = \frac{1}{K^{1/2}} \sum_{k=1}^K \left[ \frac{\hat{\sigma}_k^2 - \hat{\sigma}_{U,k}^2}{S_k} \right], \quad (2)$$

where

$$S_k^2 = \left[ (\hat{\mu}_{4,k} - \hat{\sigma}_k^4) + \hat{\delta}_{4,k}/M^4 + 2\sqrt{(\hat{\mu}_{4,k} - \hat{\sigma}_k^4)\hat{\delta}_{4,k}/M^4} \right] / J_k \quad (3)$$

and

$$\hat{\mu}_{4,k} = \sum_{j=1}^{J_k} (Y_j^{(k)} - \bar{Y}^{(k)})^4 / J_k; \quad \hat{\delta}_{4,k} = \sum_{i=1}^M \hat{p}_i^{(k)}(1 - \hat{p}_i^{(k)})(1 - 2\hat{p}_i^{(k)})^2.$$

Similarly, using items of AT2, the two variance estimates  $\hat{\sigma}_k^2$ ,  $\hat{\sigma}_{U,k}^2$  and the standard error of estimate  $S_k$  are computed and normalized within each subgroup to arrive at the statistic  $T_B$  using formula (2). The statistic  $T$  to assess departure from essential unidimensionality is given by

$$T = (T_L - T_B)/\sqrt{2}. \quad (4)$$

The null hypothesis of  $d_E = 1$  is rejected if  $T \geq Z_\alpha$ , where  $Z_\alpha$  is the upper  $100(1-\alpha)$  percentile of the standard normal distribution, and  $\alpha$  being the desired level of significance.

#### Correction for Bias in the Statistic $T_L$ by Introduction of $T_B$

Consider the statistical bias that would result if  $T_L$  rather than  $T$  were the statistic used to assess essential dimensionality. The above description shows that Stout's test is based on two variance estimates: the usual variance estimate  $\hat{\sigma}_k^2$ , and the unidimensional variance estimate  $\hat{\sigma}_{U,k}^2$ . If the items of the test measure one dominant trait, then the two subtests AT1 and PT would contain essentially unidimensional items representing the same dominant trait. When the test length is both long and essentially unidimensional,

examinees within each subgroup can be assumed to be of approximately equal ability. Consequently, it can be shown that the differences in the variance estimates  $\hat{\sigma}_k^2 - \hat{\sigma}_{U,k}^2$  computed using items in AT1, would be "small"; thus, using  $T_L$ , the test will be assessed as essentially unidimensional. By contrast, if the test length is long and essentially multidimensional, the trait measured by items of AT1 would be different from the trait(s) measured by the rest of the test, and the AT1 differences  $\hat{\sigma}_k^2 - \hat{\sigma}_{U,k}^2$  would not be small (see Stout, (1987) for the heuristics explaining why this holds), and  $T_L$  would conclude the test to be essentially multidimensional.

In the case of a relatively short essentially unidimensional test, however, examinees within each subgroup are not likely to be approximately equal on the dominant trait measured by the test, thereby causing the differences  $\hat{\sigma}_k^2 - \hat{\sigma}_{U,k}^2$  to be large. This improperly inflates the value of the statistic  $T_L$  and results in statistical bias. This bias is amplified if items of AT1 are homogeneous with respect to item difficulty, which often occurs when AT1 is selected by factor analysis. To correct for this preasymptotic bias in  $T_L$ , AT2 is constructed so that items of AT1 and AT2 are closely matched in their item difficulty distribution. It has been observed that subtests AT1 and AT2 are both subject to similar amounts of pre-asymptotic bias, but because AT2 is chosen to be similar to AT1 in difficulty only,  $T_B$  formed from AT2 will not be made larger by the presence of multidimensionality. Thus, as statistical experimental design ideas suggest, the bias is cancelled by forming the difference statistic  $T$  (Step 6, Stout, 1987).<sup>2</sup>

#### Avoiding Bias due to Guessing and High Discrimination of Items

Test items usually differ with respect to their various measurement properties. There may be difficult items, easy items, high discrimination items, low discrimination items, and so on. The 80-item SAT-Verbal vocabulary test analyzed by Lord (1968) is no

exception. Item parameter estimates for this test were obtained by LOGIST. DIMTEST with a specified level of significance  $\alpha=.05$  was applied to a three-parameter logistic, unidimensional, simulation model on various random subtests of 50 items selected from this test. For 100 replications of the DIMTEST on the simulated test data, 5 rejections of the hypothesis of  $d_E=1$  were observed—strongly confirming the unidimensional nature of the simulated data. Items of the SAT-Verbal test were then divided into two sets. One set consisted of items having discrimination parameters greater than 1.0 (high-discriminations); other set consisted of items with discrimination parameters less than or equal to 1.0 (low-discriminations). DIMTEST was applied separately to each subtest and the results were markedly different. Note from the classical test theory perspective that the first test has high reliability and the second test low reliability.

---

Table 1

---

Table 1 displays the performance of the procedure, for both subtests, administered to 750, 1000, and 2000 examinees. In these simulations, seven items were selected in each of the assessment subtests based on factor analysis with a  $J_{min}=20$ . The reported values in Table 1 are the number of rejections out of the 100 replications of DIMTEST.

The number of rejections for the test with low-discriminations is what is to be expected on a unidimensional test. However, the rejection rate for the test with high-discriminations far exceeds the nominal level of 5/100. Furthermore, as the number of examinees increases, the rejection rate also increases.

This finding was confirmed in another unidimensional simulation, which used the ASVAB general science test as its basis. Item parameter estimates for this test were

obtained by Mislevy and Bock (1984). In this simulation a rejection rate of 13/100 was observed with  $\alpha=.05$ . Further investigation showed that this elevated rejection rate was caused by a preponderance of difficult, highly discriminating items. Thus, there is evidence to show that if many items of a test are both highly discriminating and difficult with guessing present, the observed type-I error rate may be unacceptably inflated.

In an attempt to determine the cause(s) for excess bias, Monte Carlo simulations were investigated extensively with tests of high-discriminating items. Recalling that items for AT1 were chosen according to the magnitude of their loadings on the second extracted factor (Step 1, Stout, 1987), it was found that in the case of high-discriminations with guessing present (with  $d_E=1$ ), the second factor was a very pronounced *difficulty* factor even though tetrachoric correlations were used. One of the characteristics of the difficulty factor is that very easy and very difficult items have high loadings of the opposite sign. In the case of high-discriminations, for unknown reasons, but likely due to the presence of guessing, most often very easy items tended to have larger factor loadings in magnitude on the second factor than the corresponding collection of very difficult items. Consequently, the easiest items tended to be selected for AT1. To control for statistical bias, DIMTEST then selects the easiest remaining items for AT2. Therefore, PT is left with mostly difficult items. Because examinees are grouped according to their scores on PT, which mostly consists mostly of difficult items in this case, the partitioning subtest (PT) tends to misclassify low ability examinees. This misclassification is made worse if guessing is allowed. Thus, examinee abilities within each assigned subgroup may vary considerably, leading to a serious violation of the fundamental assumption of essential independence within subgroups. This assumption is critical for the statistic  $T$  to adhere closely to the nominal level of significance. As a result, the values of the statistic  $T_L$  (computed from AT1) averaged around 10, the values of  $T_B$  (computed from AT2) averaged around 7. Thus, the values of  $T = (T_L - T_B)/\sqrt{2}$  were so large that the hypothesis of  $d_E = 1$  was

often rejected.<sup>3</sup> Although  $T_B$  is supposed to compensate for the bias in  $T_L$ , the bias in  $T_L$  was so large that compensation was ineffective.

It is interesting to note that there are two reasons why the SAT subtest with low-discriminations failed to exhibit statistical bias. First, low-discriminations enhance the ability of AT2 to compensate (in a statistical, experimental design balancing sense<sup>4</sup>) for the bias contributed by items of AT1. Second, the SAT subtest with low-discriminations has a wider distribution of item difficulty, thereby tending to reduce the misclassification of examinees in the formation of subgroups.

Another unidimensional simulation study was conducted with the same high-discriminations SAT items, but with all  $c$ -parameters set to zero, creating a high-discriminations 2PL model. There was 1 rejection out of 100 trials. Therefore, the presence of guessing coupled with high-discriminations seemed to have caused the inflated rejection rate. This is true because, without guessing in the model, a highly pronounced difficulty factor is unlikely to appear in the tetrachoric factor analysis and, in fact, did not appear in high-discriminations 2PL simulations. Moreover, eliminating guessing reduces the problem of misclassification of low-ability examinees.

Based on the above findings, it was conjectured that when guessing/high discrimination items are present, the assignment of examinees to subgroups could be done more effectively using PT scores that were based on items that included easy items. This was achieved in the following way. First, items of AT1 are checked statistically, using the Wilcoxon rank sum test, to test if the items of AT1 are too easy as a group. If the Wilcoxon rank test rejects, the procedure is to replace these items with items of highest loadings of the opposite sign so that they are still dimensionally homogeneous<sup>5,6</sup>. If the Wilcoxon rank test does not reject, items of AT1 are retained. Algorithm 1 in the Appendix describes this procedure in detail. Items of AT2 are selected, as before, so that items in AT1 and AT2 have approximately the same difficulty distribution.<sup>6</sup>

### Automating the Size $M$ of Assessment Subtests

As described previously, DIMTEST splits  $N$  items of the test into three subtests: AT1 and AT2 of length  $M$  each, and PT of length  $n$  ( $= N - 2M$ ). In all the simulation studies presented in Stout (1987), the size of the assessment subtests  $M$  was specified by the user *a priori*. For example, for a 30-item test, 5 or 7 items were used in each of the assessment subtests; for a 50-item test, 8 or 12 items were used. By contrast, our aim has been to develop an algorithm that automatically determines the size of assessment subtests according to the magnitude of item loadings on the second extracted factor. For most applications this would seem preferable to the selection of  $M$ , *a priori*, especially by a novice user.

According to Stout's large sample theory for DIMTEST,  $M$  should be small compared to  $N$ . Extensive Monte Carlo simulations showed that a minimum of four items was needed in each of the assessment subtests in order to have reliable variance estimates (Nandakumar, 1987; Stout, 1984, p. 31). To determine the maximum size of  $M$  ( $\text{Max } M$ ) that will yield desirable results, three different sizes of  $M$  were tried:  $\text{Max } M = 1/5$  of the test length,  $\text{Max } M = 1/4$  of the test length, and  $\text{Max } M = 1/3$  of the test length. Similarly, to determine the minimum size of factor loading that should be used for assigning an item to AT1, three different "starting" values ( $Start$ ) of factor loadings were tried:  $Start = .25$ ,  $Start = .20$ , and  $Start = .15$ . An experimental design was set up for conducting simulations with all three sizes of  $\text{Max } M$  and with all three values of  $Start$ . For each combination of  $\text{Max } M$  and  $Start$ , both type-I error and power were observed over repeated trials of DIMTEST with tests of different types. To illustrate, let  $\text{Max } M = 1/5$  and  $Start = 0.25$ . Based on the loadings of the second factor, items with absolute loading greater than .25 are to be considered for AT1 selection. The average item loading is computed for items with positive loadings and for items with negative loadings. The set

with the highest average loading, in absolute value, is selected for AT1 and the size of this set determines  $M$ . If the minimum required number of items is not obtained with either positive or negative loadings, the start value is decreased by .05 until the minimum number of items is found. Similarly, if in the selected set more than  $1/5$  of the items have absolute loadings greater than .25, only  $1/5$  of the items with the highest loadings are included. Algorithm 2 in the Appendix describes this procedure in detail. The observation of type-I error and power for different values of  $\text{Max } M$  and  $\text{Start}$  revealed that  $\text{Max } M = 1/4$  and  $\text{Start} = .15$  yielded the most desirable results. These values were then used for selection of items in simulations reported in the Tables 3 through 7 of this paper. Other combinations of  $\text{Max } M$  and  $\text{Start}$  yielded either an observed type-I error rate that is too high or an observed power level that is too low.

#### Standard Error Estimation in Stout's Statistic

The general approach used in the development of Stout's statistic first derived an asymptotically valid test statistic and then made adjustments to optimize the pre-asymptotic behavior of the statistic, guided by Monte Carlo simulations.

Stout's statistic to test the hypothesis of essential unidimensionality was built by combining information measuring the strength of evidence of the nonunidimensionality contributed by each of the  $k = 1, \dots, K$  subgroups of examinees. That is, the goal was to construct a statistic using the quantities

$$X_k = \hat{\sigma}_k^2 - \hat{\sigma}_{U,k}^2 \quad (5)$$

from  $k$  subgroups of examinees. Each  $X_k$  measures nonunidimensionality in the sense that  $X_k \doteq 0$  when  $d_E = 1$ , and  $X_k > 0$  on average when  $d_E > 1$ . The most obvious approach is to

add up the contributions of  $X_k$  and then normalize this sum by an appropriate standard error of estimate. When unidimensionality holds, Stout (1987) found the estimated asymptotic variance of  $X_k$  to be

$$(S'_k)^2 \equiv [(\hat{\mu}_{4,k} - \hat{\sigma}_k^4) - \frac{\hat{\delta}_{4,k}}{M^4}] / J_k \quad (6)$$

leading to the statistic

$$T'' = (T'_L - T'_B) / \sqrt{2}$$

where

$$T'_L = \frac{\sum_{k=1}^K X_k}{[\sum_{k=1}^K (S'_k)^2]^{1/2}} \quad (7)$$

Result 6.1 of Stout (1987, page 599) suggests<sup>7</sup> that under regularity conditions when  $d_E = 1$ ,  $T'_L$  and  $T''$  should be asymptotically  $N(0,1)$  as the number of examinees and the number of items both approach  $\infty$ . Moreover, Result 6.4 of Stout (1987, page 601) states that both  $T'_L$  and  $T''$  should have asymptotic power one when  $d_E > 1$ .

Simulation studies conducted prior to the study reported in Stout (1987) showed that, for test lengths and examinee population sizes typically encountered in practice, the statistical test  $T''$  falsely rejected the hypothesis of unidimensionality more frequently than the nominal error rate<sup>8</sup>. Two modifications for constructing  $T$  of (4) were then considered: (a) enlarge  $S'_k$  to  $S_k$  of (3) so that the values of  $T$  on the average would be smaller, thereby reducing the rate of occurrence of type-I error to close to or even below the nominal level, and (b) normalize each  $X_k$  by its estimated standard error and then sum (instead of first summing and then normalizing as in (7)). This modified statistic  $T$  was

used in simulation studies reported in Stout (1987). However, the observed average type-I error (.023) in Stout (1987, Table 2) was well below the nominal level ( $\alpha = .05$ ).

Because  $S_k'$  yielded too large an observed type-I error and  $S_k$  yielded too small an observed type-I error, the following adjustment to the estimated standard error was considered in addition to  $S_k$  in the present study.

$$S_k'^2 = \left[ (\hat{\mu}_{4,k} - \hat{\sigma}_k^4) + \hat{\delta}_{4,k}/M^4 \right] / J_k \quad (8)$$

It can be seen that

$$S_k' \leq S_k \leq S_k$$

Furthermore, a basic question in constructing the statistic T was how to combine the building blocks  $X_k$  of (5) into a single appropriately normalized statistic for testing for unidimensionality. That is, restricting attention to linear scoring, the search was for an appropriate choice of weights  $\{a_k \mid 1 \leq k \leq K\}$  to form  $\sum_{k=1}^K a_k X_k$ . Three different weighting procedures were considered. Six new statistics  $T_1$  through  $T_6$ , as described below, were derived as a result of using different weights and standard errors of estimates. The objective was to find an improved statistic with an increased observed type-I error to approximate the nominal level while maintaining or even improving the power.

An estimator or test statistic is useful provided it centers on the appropriate parameter and had a small standard error. It can be shown that  $\text{Var}(\sum_{k=1}^K a_k X_k)$  is minimized, subject to the constraint  $\sum_{k=1}^K a_k = 1$ , by setting  $a_k = [1/\text{var}(X_k)]/\sum_{k=1}^K [1/\text{var}(X_k)]$ . Based on this argument, the statistic  $T_1 = (T_{L,1} - T_{B,1})/\sqrt{2}$  was constructed where

$$T_{L,1} = \left[ \sum_{k=1}^K \frac{X_k}{S_k^2} \right] / (\sum_{k=1}^K \frac{1}{S_k^2})^{1/2}. \quad (9)$$

The statistic  $T_2$  was constructed similar to the statistic  $T$  of (4) but with  $S'_k$  as the estimated standard error. That is,  $T_{L,2}$  is given by

$$T_{L,2} = \frac{1}{K^{1/2}} \left[ \sum_{k=1}^K \frac{X_k}{S'_k} \right]. \quad (10)$$

The statistic  $T_3$  was constructed with weights as in  $T_1$  but with  $S'_k$  as the estimated standard error. That is,  $T_{L,3}$  is given by

$$T_{L,3} = \left[ \sum_{k=1}^K \frac{X_k}{(S'_k)^2} \right] / (\sum_{k=1}^K \frac{1}{(S'_k)^2})^{1/2}. \quad (11)$$

Based upon the naive, intuitive idea that those subgroups with more examinees in them should receive more weight in the constructed statistic, two more definitions  $T_4$  and  $T_5$  were proposed, where

$$T_{L,4} = \left[ \sum_k \frac{J_k X_k}{S_k} \right] / (\sum_k J_k)^{1/2} \quad (12)$$

and

$$T_{L,5} = \left[ \sum_k \frac{\sqrt{J_k} X_k}{S_k} \right] / (\sum_k J_k)^{1/2}, \quad (13)$$

respectively.

Lastly, based upon Central Limit Theorem and contrasted with the statistic  $T$  of (3), the statistic  $T_6$  was derived where

$$T_{L,6} = [\Sigma_k X_k]/(\Sigma_k S_k^2)^{1/2}. \quad (14)$$

In summary, Stout's (1987) recommended statistics  $T$  as well as statistics  $T_1$  and  $T_6$  use  $S_k$  as the estimated standard error, and the statistics,  $T_2$  and  $T_3$  use  $S'_k$  as the estimated standard error. The statistics  $T_1$  and  $T_3$  use weights according to the principle of minimum variance with  $S_k$  and  $S'_k$  as the standard errors of estimates, respectively. The statistics  $T_4$  and  $T_5$  use weights  $J_k$  (the number of examinees in each cell) and  $\sqrt{J_k}$  respectively, with  $S_k$  as the standard error of estimate. And finally the statistic  $T_6$  is based on the usual form of the Central Limit Theorem.<sup>9</sup>

We decided that statistics  $T_i = (T_{L,i} - T_{B,i})/\sqrt{2}$ ,  $i = 1, \dots, 6$  with different weights and standard errors should provide an ample choice of statistics for a simulation study to assess whether an improved statistic can be obtained that would be better than using  $T$ .

### Monte Carlo Simulation Studies

A Monte Carlo simulation study was undertaken to study the performance of DIMTEST after performing corrections for high-discriminations bias using the Wilcoxon rank sum test, automation of the size  $M$  of assessment subtests, and correction for the standard error of estimate. In all simulations,  $J_{\min} = 20$  was adopted. The simulation study was designed to be similar to Stout's (1987) study in order to compare the performance of the statistic before and after the proposed corrections.

Two issues were of particular importance in the study: (a) how well the nominal level of significance specified by the user ( $\alpha=.05$ ) is approximated by the observed level of significance when  $d_E=1^{10}$ , and (b) how large the power of the statistical test was in various  $d_E=2$  settings.

### The Preliminary Standard Error Study

In a preliminary pilot simulation study, the performance of six different statistics  $T_1$  through  $T_6$  was studied and compared with  $T$  (after implementing corrections for high-discrimination with guessing and automated  $M$ ) with respect to type-I error and power in various test settings. The results revealed that the statistic  $T_2$  yielded a higher observed type-I error, closer to the nominal level, and a higher power than  $T$ . The statistic  $T_3$  yielded an unacceptably large type-I error; statistics  $T_1$ ,  $T_4$ ,  $T_5$ , and  $T_6$  differed little in performance from  $T$  and thus would offer no advantage. Therefore, the statistic  $T_2$  was used in the simulations described below, and the results were compared with simulations of Stout (1987), obtained by using the statistic  $T$  prior to the proposed corrections. That is,  $T$  used  $S_k$  and does not correct for high-discriminating/guessing items, nor did it automatically select  $M$ . By contrast,  $T_2$  used  $S_k$  corrected for high-discriminating/guessing items and automatically selected  $M$ .

### The Unidimensional Simulation Study

The unidimensional, three-parameter logistic model was used to simulate the test data. In order for the simulated test data to reflect real data, item parameter estimates were obtained from real data sets for five different tests: SATV, ACTM, ACTE, ASVAB AS, ASVAB AR<sup>11</sup>. The distributions of item parameters for these five tests are given in Table 2, and show that the five tests differ not only in length but also in distribution of difficulty and discrimination parameters. For example, ACTE has the lowest mean and standard deviation of item discrimination parameters; ASVAB AR had the highest mean item discrimination; ASVAB AS had the highest standard deviation of item discrimination; etc. For each test type, two examinee sample sizes  $J$  were studied: 750 and 2000. With the

sample size of 750, 250 examinees were used for factor analytic selection of assessment items, while the remainder were used to compute the test statistic. With  $J = 2000$ , 500 examinees were used for the factor analysis and the remainder were used for computing the statistic.

---

Table 2

---

Binary item responses were generated as explained below. Examinee abilities were randomly generated from the standard normal distribution. For each simulated examinee, the probability,  $P_i(\theta)$ , of correctly answering each item was computed using the three-parameter unidimensional logistic model. If a uniform random deviate in the interval  $(0,1)$  was less than or equal to the computed probability  $P_i(\theta)$ , the examinee was considered to have answered the item correctly and was given a score of 1; otherwise, the examinee was given a score of 0.

For each combination of test type and examinee size, DIMTEST (as here modified by the Wilcoxon rank sum test, automated  $M$ , and the alternate standard error of estimate  $S_k'$ ) was replicated 100 times, with new examinee responses being simulated each time. The number of rejections out of 100 replications of testing the null hypothesis of essential unidimensionality is reported in Table 3. Because the test data is generated from a unidimensional model, the observed level of significance should be close to the nominal level, which was set to .05.

---

Table 3 and Table 4

---

Table 3 shows the observed type-I error for all five simulated test types for different sample sizes. Of particular interest is the second column: rejection rates for the SATV high-discriminations. Contrasting these results with the rejection rates of Table 1 shows that, with the proposed correction for excess bias (that is, the Wilcoxon rank sum test), the rejection rates have dropped to an acceptable level. For example, the rejection rate with 2000 examinees has dropped from 58 to 7. For other test types, the observed level of significance is also close to the nominal level. Table 4 compares these results with those of Stout (1987, Table 2) where the statistic  $T$  was used. The contents of Table 4 show that, as a consequence of the proposed refinements, the observed type-I error rate has increased or remained the same for all test types and sample sizes except for ASVAB AR with 2000 examinees. The overall average observed type-I error has increased from .023 (Stout, 1987) to .045 and is very close to the nominal value of .05. In addition, for each one of the cell entries, there is no statistical evidence to reject the hypothesis that the nominal level of significance of .05 holds. That is, they are all consistent with a  $p$ -value of .05<sup>12</sup>.

#### The Two-Dimensional Simulation Study

The two-dimensional simulation study was modeled according to the multidimensional three-parameter logistic model with compensatory abilities (Reckase & McKinley, 1983) given by:

$$P_i(\theta_1, \theta_2) = c_i + \frac{1-c_i}{1 + \exp\{-1.7[a_{1i}(\theta_1-b_{1i})+a_{2i}(\theta_2-b_{2i})]\}} \quad (15)$$

Seven different test types were considered to study the power of the procedure after the proposed changes. Two-dimensional counterparts of the five test types used in the unidimensional simulation study were simulated in the following manner. The discrimination parameters ( $a_{1i}$ ,  $a_{2i}$ ) of the two dimensions for each item were independently generated from a normal distribution:

$$a_{1i} \sim N\left[\frac{\mu}{2}, \frac{\sigma}{\sqrt{2}}\right], \quad a_{2i} \sim N\left[\frac{\mu}{2}, \frac{\sigma}{\sqrt{2}}\right],$$

where  $\mu$  and  $\sigma$  were the mean and the standard deviation of the distribution of discrimination parameters of the respective unidimensional test taken from Table 2. Likewise,  $b_{1i}$  and  $b_{2i}$  were assumed to be independent of each other for each item and were generated:

$$b_{1i} \sim N(\mu, \sigma), \quad b_{2i} \sim N(\mu, \sigma),$$

where  $\mu$  and  $\sigma$  were the mean and the standard deviation of the distribution of difficulty parameters of the respective unidimensional test taken from Table 2. For example, to generate the two-dimensional counterpart of the SATV test, the  $a_{1i}$ 's and  $a_{2i}$ 's were generated independently from the normal distribution with mean  $1.07/2$  and standard deviation  $.4/\sqrt{2}$ . Similarly, the  $b_{1i}$ 's and  $b_{2i}$ 's were generated independently from the normal distribution with mean  $.58$  and standard deviation  $.88$ . Each test was taken to consist of  $N_1$  "pure" items dependent on  $\theta_1$  alone,  $N_2$  "pure" items dependent on  $\theta_2$  alone, and  $N_3$  mixed items dependent on  $\theta_1$  and  $\theta_2$ .

Abilities  $\Theta = (\Theta_1, \Theta_2)$  were generated from a bivariate normal distribution with both means being zero and both variances being one. The correlation coefficient  $\rho$  between the abilities varied appropriately. The  $c$ -parameter was taken to be .20 for all items. Binary item responses were generated exactly as described for unidimensional tests using (15).

In addition to the five two-dimensional counterparts of unidimensional tests, two more tests, the ACT Mathematics Usage Form 8B (ACTM8B) and the ACT Mathematics Usage Form 24B (ACTM24B) were used. For these two tests, estimated two-dimensional item parameters  $(a_{1i}, a_{2i})$  and  $(b_{1i}, b_{2i})$  were obtained from the American College Testing Program. Except for item parameter generation, which has been replaced by use of actual item parameter estimates, the responses for these two tests were simulated as described above.

For each of the seven test types, two examinee sample sizes  $J$  were considered—750 and 2000—and two levels of correlation  $\rho$  were considered—.5 and .7. As in the unidimensional study, when  $J=750$ , 250 examinees were used for factor analysis, and, when  $J=2000$ , 500 examinees were used for factor analysis. For each combination of test type, examinee sample size, and level of correlation, DIMTEST (as modified by the Wilcoxon rank sum test, automated  $M$ , and the alternate standard error of estimate  $S_k'$ ) was replicated 100 times, each time simulating new examinees. For the first five test types, a new set of item parameters was generated for each test after each 10 replications. The number of rejections over 100 replications is reported in Table 5 for each case.

---

Table 5 and Table 6

---

In the case of  $d_E=2$ , one wants good power; that is, one wants  $[T_2 > Z_\alpha]$  to be large for a broad range of realistic  $d_E=2$  alternatives. The contents of Table 5 show that the power is extremely high for the case of  $\rho=.5$  for both sample sizes. The power is very high for  $\rho=.7$  with 2000 examinees, and the power is good for  $\rho=.7$  with 750 examinees. These results are noteworthy, considering that all tests in the simulation study consist of at least one-third mixed items requiring knowledge of both traits to be answered correctly. Furthermore, it can be seen that as the sample size increases, the power also increases.

Table 6 compares the results of the present study with the results of Stout's simulation study which uses the statistic  $T$  (1987, Table 6). It can be seen, as a consequence of the proposed refinements, that the power has increased for every test type, sample size, and level of correlation. On the average, power has gone up from 67 to 88 rejections per 100 trials of the procedure for the case of  $\rho=.5$  with 750 examinees, from 92 to 99 rejections for the case of  $\rho=.5$  with 2000 examinees, from 36 to 54 for the case of  $\rho=.7$  with 750 examinees, and from 67 to 90 rejections for the case of  $\rho=.7$  with 2000 examinees. These average increases are large enough to be of practical importance.

### Real Data Study

Four different data sets were used to examine the performance of DIMTEST on actual data. Data for two Armed Services Vocational Aptitude Batteries, used by the Department of Defense Student Testing Program in high schools and post-secondary schools, were obtained from Linn, Hastings, Hu, and Ryan (1987). These tests included Arithmetic Reasoning tests for Grades 10 and 12 (AR10 & AR12), each with 30 items and 1984 and 1961 examinees, respectively. Two more data sets were obtained from American College Testing (ACT) Program. These included ACT mathematics usage Forms B and C (F29B & F29C), each with 40 items and 2491 and 2494 examinees, respectively.

DIMTEST was applied to each of the four data sets. In each data set, 500 examinees were randomly selected for factor analysis; the rest were used for computing the statistic. Examinees were randomly split into two groups, one group for performing factor analysis and the other for computing the statistic, 100 times—each time testing for the null hypothesis of essential unidimensionality. The number of rejections over 100 replications of the procedure noted. The results for all tests are tabulated in Table 7.

---

Table 7

---

The contents of Table 7 suggest that, according to the DIMTEST, AR10 and AR12 should be assessed as essentially unidimensional tests while F29B and F29C should be assessed as multidimensional tests. Examination of items of F29B and F29C showed that these tests consist of items assessing knowledge of arithmetic and algebra operations, geometry, numeration, story problems, and advanced topics. Therefore, from the perspective of content, F29B and F29C would seem to be multidimensional tests measuring highly correlated abilities. The rejection rate for AR12 is slightly higher than expected for an essentially unidimensional test. One or two items highly influenced by another factor may contribute to this high rejection rate, or many items may be slightly influenced by a second factor. Further investigation is necessary to examine possible reasons.

#### Summary and Discussion

Detailed investigation of DIMTEST for assessing unidimensionality revealed certain limitations. It failed to perform desirably when the test consisted of predominantly

difficult, high-discrimination items coupled with guessing present. This limitation was overcome by a more appropriate selection of assessment items. Also, an automated approach was devised to determine the size of assessment subtests, and the estimate of the standard error of the statistic was adjusted to yield the desired level of significance for  $d_E=1$  data and higher power for  $d_E>1$  data. After the proposed refinements were implemented, DIMTEST was applied to a variety of simulated tests for different sample sizes; these tests were modeled on Stout's (1987) simulation study.

Comparison of the results of the present study with the results of Stout's (1987) study indicates that the proposed refinements have improved the observed level of significance. It is now close to the nominal level for  $d_E=1$  simulations and has considerably increased the power for  $d_E=2$  simulations for different levels of correlations and sample sizes. In addition, the procedure has been used on a number of real data sets. The results of the real test data study seem to confirm the a priori hypotheses regarding the dimensionality of these tests.<sup>13</sup>

The refinements have led to a revised test procedure that is, in particular, more robust against unusually high-discrimination parameters with guessing present and that, in general, is able to perform more desirably with respect to type-I and type-II errors. Moreover, the procedure is automated and totally data-dependent in its selection of assessment subtest items, making it more user friendly. The automation of the size of the assessment subtests could especially benefit the novice user. Because the power of the statistical test heavily relies upon appropriate selection of items for AT1, our simulation study provides further evidence that the use of linear factor analysis for selection of these items is a promising approach that requires little effort on the part of the user.

When the statistical test rejects the null hypothesis of essential unidimensionality, it is possible to proceed in several ways. One approach would be to reexamine the test and assess the complexity of the essential multidimensionality present using DIMTEST,

NOFA, and so forth. If inference suggests that each of the different dominant traits influences a distinct group of items (i.e., there is a pronounced simple structure), the test could be split into several essentially unidimensional subtests, and each one could be analyzed separately using unidimensional IRT models. Alternately, if most of the items of the test are each influenced simultaneously by several dominant dimensions, then the researcher may need to resort to multidimensional parametric models in order to make inferences about the test data (Reckase, 1985, 1989).

The dimensionality of a set of item responses is conceptually very complex. It is a function of items, examinees, and extraneous factors such as type of instruction and stage of learning. Also, dimensionality is, from the practical perspective, a continuum. Because items are multiply determined, among finite length tests (the only kind available in applications), there is no such thing as a strictly unidimensional test. But we can still describe a given set of item responses as being well modeled by an essentially unidimensional test model. Junker (1990, 1991) argues that an index for the continuum of dimensionality should be developed with strict unidimensionality, in the sense of fitting local independence models on one end and strict essential multidimensionality on the other end, with essential unidimensionality in between. Junker and Stout (1991) have developed indices for lack of essential unidimensionality, which can be extremely useful for assessing the degree of lack of essential unidimensionality when Stout's test of  $d_E=1$  is rejected. Additionally, these indices show when it is safe to use unidimensional estimation procedures such as LOGIST or BILOG to arrive at accurate ability estimates. The conjecture is that lack of strict unidimensionality is not detrimental, provided  $d_E=1$  modeling provides a good approximation to reality. The number of items influenced by the secondary dimensions, as well as the strength of the influence of secondary dimensions, on each item should determine how strong the lack of  $d_E=1$  is. Nandakumar (1991) has demonstrated the utility of DIMTEST in assessing essential unidimensionality when test

items were influenced by various dimensions to various degrees, and thus strict dimensionality exceeded one. Nandakumar has found that the accuracy of the approximation of essential unidimensionality for a test is a function of the proportion of test items influenced by the various nondominant traits present and by the strength of the influence of these traits.

Stout's procedure seems very promising for assessing the dimensionality underlying a set of items. It is an outgrowth of the conceptual definition of essential unidimensionality and was developed to be sensitive to dominant dimensions and insensitive to transient or minor dimensions. The procedure is nonparametric (thus avoiding parametric model-data fit problems), supported by an asymptotic theory, and is computationally simplistic. However, the procedure is relatively new, and its applicability in a variety of realistic applications needs to be studied further. Software to run DIMTEST is available from the authors.

## Appendix

### Algorithm 1: Test for Difficulty Factor

1. Rank the  $N$  items from most difficult (rank 1) to easiest (rank  $N$ ).
2. Compute the sum  $W_s$  of the ranks of the  $M$  items in AT1.
3. Compute the mean  $E(W_s)$  and the standard deviation  $SD(W_s)$  of the sum  $W_s$  under the assumption of randomly distributed ranks:

$$E(W_s) = \frac{1}{2} M(N+1)$$

$$SD(W_s) = \left( \frac{1}{12} M(N-M)(N+1) \right)^{1/2}.$$

4. Compute the critical value  $C$  for  $W_s$  under the usual large sample approximation:

$$C = E(W_s) + Z_\alpha(SD(W_s)),$$

where  $Z_\alpha$  is the upper  $100(1-\alpha)$ th percentile of the standard normal distribution and  $\alpha$  is the desired level of significance.

5. If  $W_s > C$ , conclude that  $M$  items in AT1 are too easy.

**Algorithm 2: The Size M of Assessment Subtests**

Let  $N$  = total number of items,  $Mlow = 4$ ,  $Mhigh = \left\lceil \frac{N}{4} \right\rceil$ , and  $Maxload = .15$ .

## 1. Compute

- a)  $I_1$  = Number of positive loadings  $\geq Maxload$ .
- b)  $I_2$  = Number of negative loadings  $\leq -Maxload$ .

## 2. Redefine

$$\begin{aligned} I_1 &:= \min(Mhigh, I_1) \\ I_2 &:= \min(Mhigh, I_2). \end{aligned}$$

3. If both  $I_1 < Mlow$  and  $I_2 < Mlow$ , then define

$$Maxload := Maxload - .05.$$

Go to Step 1.

4. If either  $I_1$  or  $I_2$  is  $\geq Mlow$ , then let

$$M = \begin{cases} I_1 & \text{if } I_1 \geq Mlow \\ I_2 & \text{if } I_2 \geq Mlow \end{cases}$$

5. If both  $I_1 \geq Mlow$  and  $I_2 \geq Mlow$ , then compute the averages Avg1 and Avg2 of item loadings for sets corresponding to  $I_1$  and  $I_2$  respectively. Let

$$M = \begin{cases} I_1 & \text{if } \text{Avg1} > \text{Avg2} \\ I_2 & \text{if } \text{Avg2} > \text{Avg1} \\ \text{Max}(I_1, I_2) & \text{if } \text{Avg1} = \text{Avg2} \end{cases}$$

Notes

<sup>1</sup>Throughout, we speak of a unidimensional test, a unidimensional set of items, etc. This convenient phrasing represents the more complex reality that the dimensionality of a model or a data set rests on the joint influence of test items and examinee population. Items and examinees together produce test data that we judge by statistical inference to be unidimensional or not. Reckase (1990) writes perceptively on this point. Technically, IRT dimensionality is usually defined to be the lowest latent space dimension possible, such that monotonicity and local independence hold.

<sup>2</sup>Note that the statistic  $T_L$  computed from AT1 is sensitive to dimensionality (that is, it can discriminate between  $d_E=1$  vs  $d_E>1$ ) and to sources of bias. The idea in introducing AT2 is to deliberately make  $T_B$  sensitive only to sources of bias but not to dimensionality.

<sup>3</sup>In unidimensional settings where the procedure worked well, typical values of  $T_L$  ranged roughly from 1 to 5, and typical values of  $T_B$  ranged roughly between .6 to 4.0; thus typical values of T ranged roughly between -1.0 to 1.5.

<sup>4</sup>If a randomized block design with  $M$  blocks of size 2 is to be used in an experiment with human subjects assigned to control and treatment groups, this experimental design technique will work well unless the subjects are too variable. By rough analogy, the higher the discrimination parameters, the more "variable" are the items that are being assigned to AT1 and AT2 and the less effective the difficulty matching method (analogous to blocking) of AT2 item selection is in eliminating bias.

<sup>5</sup>It can be observed that when items in AT1 are replaced (because they are too easy) with

items of high loadings of the opposite sign, easy PT items could result, thereby causing inaccuracy in subgroup assignment of high-scoring examinees. Simulation results have shown that this potential inaccuracy is not as detrimental to the value of the statistic T as it was when PT had mostly difficult items.

<sup>6</sup>We also tried to correct tetrachoric correlations for guessing by following Bock, Gibbons, and Muraki (1985) and by using nonlinear factor analyses to diminish the influence of difficulty on the second factor loadings. Regarding correction of tetrachorics, we found that when guessing values were about .2 in the model, a large percentage of the sample correlations was computed as 1 or -1. However, when the guessing levels were arbitrarily cut by half, the problem of extreme correlations was reduced. Even with this reduction of guessing levels, the items selected for AT1 did not differ significantly from those selected without correction for guessing. Moreover, the ad hoc method of cutting guessing levels defeats the purpose of using the three-parameter logistic model. Therefore correction for guessing was not implemented.

The nonlinear factor extraction program NOFA was used to select items for the assessment subtests. We tried two-factor quadratic model for this purpose. In comparing the results of linear and nonlinear factor analysis, we found no difference in T-values between the two methods. To our surprise the difficulty factor reappeared even with nonlinear factor analysis. Therefore, we did not implement nonlinear factor analysis.

<sup>7</sup>The reason for the word "suggests" instead of "establishes" is that Stout's result actually assumes unidimensionality under the stronger assumption of local independence. Further, the asymptotic invariance in (6) also assumes the stronger assumption of local independence.

<sup>8</sup>( $S_k'$ )<sup>2</sup> is an asymptotic variance and fails to account for the overdispersion of  $S_k$  that occurs as examinees in a fixed PT subgroup have varying abilities, even though the test is unidimensional. Thus, ( $S_k'$ )<sup>2</sup> will underestimate the true standard error and will yield too large a type-I error (see Cox & Snell ,1989, pp 106-110 for a nice discussion of overdispersion resulting from varying parameter such as ability).

<sup>9</sup>There are, of course, many more possibilities for computing statistics with given weights and standard errors of estimate, but those described here were considered the most appropriate.

<sup>10</sup>Technically, our simulations were done with  $d=1$ , implying  $d_E=1$ . For simulation studies for which  $d_E=1$ , see Nandakumar (1991).

<sup>11</sup>The SATV denotes the SAT-verbal test obtained from Lord (1968); ACTM denotes the ACT mathematics usage test, and ACTE denotes the ACT English Usage test, both obtained from Drasgow (1987); ASVAB AS and ASVAB AR denote the Armed Services Vocational Aptitude test Battery, Auto Shop Information and Arithmetic Reasoning respectively, both obtained from Mislevy and Bock (1984).

<sup>12</sup>The standard error for testing the hypothesis of  $p=.05$  vs  $p \neq .05$  is approximately 2.2 trials. Thus, the acceptance region of this test for a set of 100 simulations is given by (.7, 9.3) trials.

<sup>13</sup>We say "seem to" because one cannot really know that a real data set is  $d_E=1$  or  $d_E>1$ . Further, the 100 replications of Table 7 are not the result of 100 administrations of the test to similar examinee populations, but rather 100 variations of the application of the statistic

to one data set that resulted from one administration of the test.

## REFERENCES

- Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Educational Measurement, 17, 283-296.
- Bejar, I. I. (1983). Introduction to item response theory models and their assumptions. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 1-23). British Columbia: Educational Research Institute of British Columbia.
- Berger, M. P., & Knol, D. L. (1990, April). On the assessment of dimensionality in multidimensional item response theory models. Paper presented at the annual meeting of the American Educational Research Association, Boston.
- Birenbaum, M., & Tatsuoka, K. K. (1982). On the dimensionality of achievement data. Journal of Educational Measurement, 19, 259-266.
- Bock, R. D., Gibbons, R., & Muraki, E. (1985). Full-information item factor analysis (MRC Report No. 85-1). Chicago: National Opinion Research Center.
- Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items and between tests. Psychometrika, 26, 347-372.
- Christoffersson, A. (1975). Factor analysis of dichotomized variables. Psychometrika, 40, 5-32.
- Cox, D. R., & Snell, E. J. (1989). Analysis of Binary Data, London: Chapman-Hall.
- Drasgow, F. (1987). A study of measurement bias of two standard psychological tests. Journal of Applied Psychology, 72, 19-30.
- Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. Applied Psychological Measurement, 10, 287-302.
- Hambleton, R. K., & Swaminathan, H. (1985). Item Response Theory: Principles and applications, Kluwer-Nijhoff Publishers, Boston.
- Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. Multivariate Behavioral Research, 19, 49-78.
- Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. Applied Psychological Measurement, 9, 139-164.
- Holland, P. W. (1981). When are item response models consistent with observed data? Psychometrika, 46, 79-92.
- Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. Annals of Statistics, 14, 1523-1543.
- Hulin, C. L., Drasgow, F., & Parsons, L. K. (1983). Item Response Theory.

- Homewood, Illinois. Dow Jones-Irwin.
- Humphreys, L. G. (1981). The primary mental ability. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds). Intelligence and learning (pp. 87-102). New York: Plenum.
- Humphreys, L. G. (1985). General intelligence: An integration of factor, test, and simplex theory. In B. B. Wolman (Ed.), Handbook of intelligence (pp. 201-224). New York: Wiley.
- Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. Journal of Applied Psychology, 71, 327-333.
- Etazadi-Amoli, J. E., & McDonald, R. P. (1983). A second generation nonlinear factor analysis. Psychometrika, 48, 315-342.
- Junker, B. (1988). Statistical aspects of a new latent trait theory, Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.
- Junker, B. (1990, June). Essential independence and structural robustness in item response theory. Paper presented at the annual meeting of the Psychometric Society, Princeton.
- Junker, B. (1991). Essential independence and likelihood-based ability estimation for polytomous items. Psychometrika, 56, 255-278.
- Junker, B., & Stout, W. (1991, July). Structural robustness of ability estimates in item response theory. Paper presented at the 7th European Meeting of the Psychometric Society, Trier, Germany.
- Linn, R. L., Hastings, N. C., Hu, G., & Ryan, K. E. (1987). Armed Services Vocational Aptitude Battery: Differential item functioning on the high school form. Dayton, OH: USAF Human Resources Laboratory.
- Lord, F. M. (1968). An analysis of verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 28, 989-1020.
- McDonald, R. P. (1962). A general approach to nonlinear factor analysis. Psychometrika, 29, 397-415.
- McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. British Journal of Mathematical and Statistical Psychology, 27, 82-89.
- Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. Journal of Educational Statistics, 11, 3-31.
- Mislevy, R. J., & Bock, R. D. (1984). Item operating characteristics of the Armed Services Aptitude Battery (ASVAB). (Tech. Rep. No N00014-83-C-0283), Chicago IL: Office of Naval Research.
- Muthen, B. (1978). Contributions to factor analysis of dichotomous variables. Psychometrika, 43, 551-560.

- Nandakumar, R. (1987). Refinements of Stout's procedure for assessing latent trait dimensionality. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.
- Nandakumar, R. (1991). Traditional dimensionality vs. essential dimensionality. Journal of Educational Measurement, 28, 99-117.
- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4, 207-230.
- Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. Applied Psychological Measurement, 9, 401-412.
- Reckase, M. D. (1989). The interpretation and application of multidimensional item response theory models; and computerized testing in the instructional environment. Iowa City, IA: American College Testing Program.
- Reckase, M. D. (1990, April). Unidimensionality data from multidimensional tests and multidimensional data from unidimensional tests. Paper presented at the annual meeting of the American Educational Research Association, Boston.
- Reckase, M. D., & McKinley, R. L. (1983, April). The definition of difficulty and discrimination for multidimensional item response theory models. Paper presented at the annual meeting of the American Educational Research Association, Montreal.
- Roznowski, M. A., Tucker, L. R., & Humphreys, L. G. (1991). Three approaches to determining the dimensionality of binary data. Applied Psychological Measurement, 15, 109-128.
- Stout, W. F. (1984). The statistical assessment of latent trait dimensionality in psychological testing. Rep. No. N00014-82-K-0486. Urbana, IL: Office of Naval Research Technical Report.
- Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. Psychometrika, 52, 589-617.
- Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. Psychometrika, 55, 293-326.
- Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), Applications of item response theory. British Columbia: Educational Research Institute of British Columbia.
- Yen, W. M. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. Psychometrika, 50, 399-410.
- Zwick, R. (1987a). Assessment of dimensionality of NEAP Year 15 reading data. (ETS Research Report 86-4.) Princeton, N.J.: Educational Testing Service.
- Zwick, R. (1987b). Assessing the dimensionality of NAEP reading data. Journal of Educational Measurement, 24, 293-308.

Table 1  
 Rejection rates per 100 trials for  $d_E=1$  simulation study using  
 estimated item parameters of SAT verbal test with  $\alpha=0.05$

Discrimination parameter	Number of items	Number of examinees		
		750	1000	2000
$0 \leq \alpha_i \leq 1.0$ (low-discriminations)	41	4	0	3
$1.1 \leq \alpha_i \leq 2.0$ (high-discriminations)	39	28	46	58

**Table 2**  
 Sample distributions of item parameters for the five  
 standardized tests used in the study

N : **	SATV * 80	ACTM 40	ACTE 75	ASVAB	
				AS	AR 30
Max $a_i$ 's	2.00	2.00	1.58	2.82	2.76
Min $a_i$ 's	0.40	0.40	0.11	0.32	0.50
Mean $a_i$ 's	1.07	1.09	0.72	1.22	1.46
S.D $a_i$ 's	0.40	0.35	0.25	0.70	0.51
Max $b_i$ 's	2.50	1.50	2.07	1.27	1.01
Min $b_i$ 's	-1.50	-1.02	-3.11	-1.39	-2.72
Mean $b_i$ 's	0.58	0.50	0.03	0.09	-0.02
S.D $b_i$ 's	0.88	0.61	0.96	0.72	0.84
Max $c_i$ 's	0.20	0.21	0.27	0.26	0.34
Min $c_i$ 's	0.04	0.02	0.04	0.06	0.08
Mean $c_i$ 's	0.16	0.14	0.15	0.20	0.19
S.D $c_i$ 's	0.05	0.04	0.03	0.04	0.06

\*\* N denotes the test length.

\* SATV denotes the SAT verbal test battery.  
 ACTM denotes the ACT mathematics usage test battery.  
 ACTE denotes the ACT English usage test battery.  
 ASVAB AS denotes the Armed Services Vocational Aptitude Battery for auto shop information.  
 ASVAB AR denotes the Armed Services Vocational Aptitude Battery for arithmetic reasoning.

Table 3

Results of unidimensional simulation study: Rejection rates for testing the null hypothesis of  $d_E=1$  over 100 trials with  $c=.20$  and  $\alpha=.05$

<i>J</i>	SATV		ACTM	ACTE	ASVAB AS	ASVAB AR
	SATV*	high dis				
750	6	8	5	6	2	3
2000	6	7	4	4	2	1

\* SATV and ACTE each contain more than 50 items in the pool, but 50 items were randomly selected for the study. After each 10 of 100 trials a new sample of 50 items was chosen. For other tests the same test was used for all 100 trials.

Table 4

Comparison of unidimensional simulation study results of this paper with those in Stout (1987): Rejection rates for testing the null hypothesis of  $d_E=1$  over 100 trials with  $c=.2$  and  $\alpha=.05$

Study	SATV		ACTM		ACTE		ASVAB AS		ASVAB AR	
	750	2000	750	2000	750	2000	750	2000	750	2000
Stout * (1987)	2	6	1	4	3	1	1	1	2	4
Present	6	6	5	4	6	4	2	2	3	1

\* For all tests the rejection rate reported is the average of rejection rates (rounded to nearest integer) for the two different M values reported in Table 2 of Stout (1987).

Table 5

Results of two-dimensional simulation study: Rejection rates for testing the null hypothesis of  $d_E=1$  over 100 trials with  $c=.20$  and  $\alpha=.05$

$N_1 - N_2 - N_3$	SATV			ACTM			ACTE			ASVAB AR		ASVAB AR		ACTM24B		ACTM8B					
	17	17	16	13	13	14	17	17	16	8	8	9	10	10	10	0	0	40	0	0	50
$J$	750	2000	750	2000	750	2000	750	2000	750	2000	750	2000	750	2000	2000	2000	2000	2000	2000	2000	
$\rho = .5$	93	100	97	100	81	100	73	99	94	98	99	100									
$\rho = .7$	58	96	66	97	37	83	50	93	61	91	69	98									

Table 6

Comparison of two-dimensional simulation study results of this paper with those in Stout (1987): Rejection rates over 100 trials for testing the null hypothesis  $d_E=1$  with  $c=.2$ , and  $\alpha=.05$

$N_1 - N_2 - N_3:$	SATV 17-17-16		ACTM 13-13-14		ACTE 17-17-16		ASVAB AS 8-8-9		ASVAB AR 10-10-10		
$J$	750	2000	750	2000	750	2000	750	2000	750	2000	
$\rho=.5$	Stout * (1987)	62	98	69	-	59	90	-	87	76	-
	Present	93	100	97	100	81	100	73	99	94	98
$\rho=.7$	Stout * (1987)	36	83	-	74	-	55	-	54	-	67
	Present	58	96	66	97	37	83	50	83	61	91

\* For all tests the rejection rate reported is the average of rejection rates (rounded to nearest integer) for the two different  $M$  values reported in Table 6 of Stout (1987).

Table 7  
Results of real data study: Rejection rates for testing  
the null hypothesis of  $d_E=1$  over 100 replications of random  
selection of subjects with  $\alpha=.05$

	AR10	AR12	F29B	F29C
<i>N:</i>	30	30	40	40
<i>J:</i>	1984	1961	2491	2494
	6	13	86	82

Distribution List

- Dr. Terry Ackerman  
Educational Psychology  
260C Education Bldg.  
University of Illinois  
Champaign, IL 61801
- Dr. Terry Allard  
Code 1143CS  
Office of Naval Research  
800 N. Quincy St.  
Arlington, VA 22217-5000
- Dr. Nancy Allen  
Educational Testing Service  
Princeton, NJ 08541
- Dr. Gregory Anrig  
Educational Testing Service  
Princeton, NJ 08541
- Dr. Phipps Arabie  
Graduate School of Management  
Rutgers University  
92 New Street  
Newark, NJ 07102-1895
- Dr. Isaac I. Bejar  
Law School Admissions  
Services  
Box 40  
Newtown, PA 18940-0040
- Dr. William O. Berry  
Director of Life and  
Environmental Sciences  
AFOSR/NL N1, Bldg. 410  
Bolling AFB, DC 20332-6448
- Dr. Thomas G. Bever  
Department of Psychology  
University of Rochester  
River Station  
Rochester, NY 14627
- Dr. Menucha Birenbaum  
Educational Testing  
Service  
Princeton, NJ 08541
- Dr. Bruce Blox  
Defense Manpower Data Center  
99 Pacific St.  
Suite 155A  
Monterey, CA 93943-3231
- Dr. Gwyneth Bodoo  
Educational Testing Service  
Princeton, NJ 08541
- Dr. Richard L. Branch  
HO, USMEPCOM/MEPCT  
2510 Green Bay Road  
North Chicago, IL 60064
- Dr. Robert Brennan  
American College Testing  
Programs  
P.O. Box 168  
Iowa City, IA 52243
- Dr. David V. Budescu  
Department of Psychology  
University of Haifa  
Mount Carmel, Haifa 31999  
ISRAEL
- Dr. Gregory Candell  
CTB/McMillan/McGraw-Hill  
2500 Garden Road  
Monterey, CA 93940
- Dr. Paul R. Chatekier  
Perceptronics  
1911 North Ft. Myer Dr.  
Suite 100  
Arlington, VA 22209
- Dr. Susan Chipman  
Cognitive Science Program  
Office of Naval Research  
800 North Quincy St.  
Arlington, VA 22217-5000
- Dr. Raymond E. Christal  
UES LAMP Science Advisor  
AL/HRMIL  
Brooks AFB, TX 78235
- Dr. Norman Cliff  
Department of Psychology  
Univ. of So. California  
Los Angeles, CA 90089-1061
- Director  
Life Sciences, Code 1142  
Office of Naval Research  
Arlington, VA 22217-5000
- Commanding Officer  
Naval Research Laboratory  
Code 4827  
Washington, DC 20375-5000
- Dr. John M. Cornwell  
Department of Psychology  
I/O Psychology Program  
Tulane University  
New Orleans, LA 70118
- Dr. William Crano  
Department of Psychology  
Texas A&M University  
College Station, TX 77843
- Dr. Linda Curran  
Defense Manpower Data Center  
Suite 400  
1600 Wilson Blvd.  
Rosslyn, VA 22209
- Dr. Timothy Davey  
American College Testing Program  
P.O. Box 168  
Iowa City, IA 52243
- Dr. Charles E. Davis  
Educational Testing Service  
Mail Stop 22-T  
Princeton, NJ 08541
- Dr. Ralph J. DeAyala  
Measurement, Statistics,  
and Evaluation  
Benjamin Bldg., Rm. 1230P  
University of Maryland  
College Park, MD 20742
- Dr. Sharon Derry  
Florida State University  
Department of Psychology  
Tallahassee, FL 32306
- Hei-Ki Dong  
Bellcore  
6 Corporate Pl.  
RM: PYA-1K207  
P.O. Box 1320  
Piscataway, NJ 08855-1320
- Dr. Neil Dorans  
Educational Testing Service  
Princeton, NJ 08541
- Dr. Fritz Drasgow  
University of Illinois  
Department of Psychology  
403 E. Daniel St.  
Champaign, IL 61820
- Defense Technical  
Information Center  
Cameron Station, Bldg. 5  
Alexandria, VA 22314  
(2 Copies)
- Dr. Richard Duran  
Graduate School of Education  
University of California  
Santa Barbara, CA 93106
- Dr. Susan Embretson  
University of Kansas  
Psychology Department  
426 Fraser  
Lawrence, KS 66045
- Dr. George Engelhardt, Jr.  
Division of Educational Studies  
Emory University  
210 Fishburne Bldg.  
Atlanta, GA 30322
- ERIC Facility-Acquisitions  
2440 Research Blvd., Suite 550  
Rockville, MD 20850-3238
- Dr. Marshall J. Farr  
Farr-Sight Co.  
2520 North Vernon Street  
Arlington, VA 22207
- Dr. Leonard Feldt  
Lindquist Center  
for Measurement  
University of Iowa  
Iowa City, IA 52242
- Dr. Richard L. Ferguson  
American College Testing  
P.O. Box 168  
Iowa City, IA 52243
- Dr. Gerhard Fischer  
Liebigasse 5  
A 1010.Vienna  
AUSTRIA
- Dr. Myron Fischl  
U.S. Army Headquarters  
DAPE-HR  
The Pentagon  
Washington, DC 20310-0300
- Mr. Paul Foley  
Navy Personnel R&D Center  
San Diego, CA 92152-6800
- Chair, Department of  
Computer Science  
George Mason University  
Fairfax, VA 22030
- Dr. Robert D. Gibbons  
University of Illinois at Chicago  
NPI 909A, M/C 913  
912 South Wood Street  
Chicago, IL 60612
- Dr. Janice Gifford  
University of Massachusetts  
School of Education  
Amherst, MA 01003
- Dr. Robert Glaser  
Learning Research  
& Development Center  
University of Pittsburgh  
3939 O'Hara Street  
Pittsburgh, PA 15260
- Dr. Susan R. Goldman  
Peabody College, Box 45  
Vanderbilt University  
Nashville, TN 37203
- Dr. Timothy Goldsmith  
Department of Psychology  
University of New Mexico  
Albuquerque, NM 87131

- Dr. Sherrie Gott  
AFHRL/MOMJ  
Brooks AFB, TX 78235-5601
- Dr. Bert Green  
Johns Hopkins University  
Department of Psychology  
Charles & 34th Street  
Baltimore, MD 21218
- Prof. Edward Haertel  
School of Education  
Stanford University  
Stanford, CA 94305-3096
- Dr. Ronald K. Hambleton  
University of Massachusetts  
Laboratory of Psychometric  
and Evaluative Research  
Hills South, Room 152  
Amherst, MA 01003
- Dr. Delwyn Hamisch  
University of Illinois  
51 Gerty Drive  
Champaign, IL 61820
- Dr. Patrick R. Harrison  
Computer Science Department  
U.S. Naval Academy  
Annapolis, MD 21402-5002
- Ms. Rebecca Hettler  
Navy Personnel R&D Center  
Code 13  
San Diego, CA 92152-6800
- Dr. Thomas M. Hirsch  
ACT  
P.O. Box 168  
Iowa City, IA 52243
- Dr. Paul W. Holland  
Educational Testing Service, 21-T  
Rosedale Road  
Princeton, NJ 08541
- Prof. Lutz F. Hornke  
Institut für Psychologie  
RWTH Aachen  
Jaegerstrasse 17/19  
D-5100 Aachen  
WEST GERMANY
- Ms. Julia S. Hough  
Cambridge University Press  
40 West 20th Street  
New York, NY 10011
- Dr. William Howell  
Chief Scientist  
AFHRL/CA  
Brooks AFB, TX 78235-5601
- Dr. Huynh Huynh  
College of Education  
Univ. of South Carolina  
Columbia, SC 29208
- Dr. Martin J. Ippel  
Center for the Study of  
Education and Instruction  
Leiden University  
P.O. Box 9555  
2300 RB Leiden  
THE NETHERLANDS
- Dr. Robert Jannarone  
Elec. and Computer Eng. Dept.  
University of South Carolina  
Columbia, SC 29208
- Dr. Kumar Joag-dev  
University of Illinois  
Department of Statistics  
101 Illini Hall  
725 South Wright Street  
Champaign, IL 61820
- Professor Douglas H. Jones  
Graduate School of Management  
Rutgers, The State University  
of New Jersey  
Newark, NJ 07102
- Dr. Brian Junker  
Carnegie-Mellon University  
Department of Statistics  
Pittsburgh, PA 15213
- Dr. Marcel Just  
Carnegie-Mellon University  
Department of Psychology  
Schenley Park  
Pittsburgh, PA 15213
- Dr. J. L. Kaiwei  
Code 442/JK  
Naval Ocean Systems Center  
San Diego, CA 92152-5000
- Dr. Michael Kaplan  
Office of Basic Research  
U.S. Army Research Institute  
5001 Eisenhower Avenue  
Alexandria, VA 22333-5600
- Dr. Jeremy Kilpatrick  
Department of  
Mathematics Education  
105 Adelheid Hall  
University of Georgia  
Athens, GA 30602
- Ms. Hae-Rim Kim  
University of Illinois  
Department of Statistics  
101 Illini Hall  
725 South Wright St.  
Champaign, IL 61820
- Dr. Jwa-keun Kim  
Department of Psychology  
Middle Tennessee State  
University  
Murfreesboro, TN 37132
- Dr. Sung-Hoon Kim  
KEDI  
92-6 Umyeon-Dong  
Seocho-Gu  
Seoul  
SOUTH KOREA
- Dr. G. Gage Kingsbury  
Portland Public Schools  
Research and Evaluation Department  
501 North Dixon Street  
P.O. Box 3107  
Portland, OR 97209-3107
- Dr. William Koch  
Box 7246, Meas. and Eval. Cir.  
University of Texas-Austin  
Austin, TX 78703
- Dr. James Kraatz  
Computer-based Education  
Research Laboratory  
University of Illinois  
Urbana, IL 61801
- Dr. Patrick Kyllonen  
AFHRL/MOEL  
Brooks AFB, TX 78235
- Ms. Carolyn Laney  
1515 Spencerville Rd  
Spencerville, MD 20868
- Richard Lanterman  
Commandant (G-PWP)  
US Coast Guard  
2100 Second St., SW  
Washington, DC 20593-0001
- Dr. Michael Levine  
Educational Psychology  
210 Education Bldg.  
1310 South Sixth Street  
University of IL at  
Urbana-Champaign  
Champaign, IL 61820-6990
- Dr. Charles Lewis  
Educational Testing Service  
Princeton, NJ 08541-0001
- Mr. Hsin-hung Li  
University of Illinois  
Department of Statistics  
101 Illini Hall  
725 South Wright St.  
Champaign, IL 61820
- Library  
Naval Training Systems Center  
12350 Research Parkway  
Orlando, FL 32826-3224
- Dr. Marcia C. Linn  
Graduate School  
of Education, EMST  
Tolman Hall  
University of California  
Berkeley, CA 94720
- Dr. Robert L. Linn  
Campus Box 249  
University of Colorado  
Boulder, CO 80309-0249
- Logicon Inc. (Attn: Library)  
Technical and Training Systems  
Division  
P.O. Box 85158  
San Diego, CA 92138-5158
- Dr. Richard Luecht  
ACT  
P.O. Box 168  
Iowa City, IA 52243
- Dr. George B. Macready  
Department of Measurement  
Statistics & Evaluation  
College of Education  
University of Maryland  
College Park, MD 20742
- Dr. Evans Mandes  
George Mason University  
4400 University Drive  
Fairfax, VA 22030
- Dr. Paul Mayberry  
Center for Naval Analysis  
4401 Ford Avenue  
P.O. Box 16268  
Alexandria, VA 22302-0118
- Dr. James R. McBride  
HumRRO  
6430 Elmhurst Drive  
San Diego, CA 92120
- Mr. Christopher McCusker  
University of Illinois  
Department of Psychology  
603 E. Daniel St.  
Champaign, IL 61820
- Dr. Robert McKinley  
Educational Testing Service  
Princeton, NJ 08541

Dr. Joseph McLachlan  
Navy Personnel Research  
and Development Center  
Code 14  
San Diego, CA 92152-6800

Alan Mead  
c/o Dr. Michael Levine  
Educational Psychology  
210 Education Bldg.  
University of Illinois  
Champaign, IL 61801

Dr. Timothy Miller  
ACT  
P.O. Box 168  
Iowa City, IA 52243

Dr. Robert Mislevy  
Educational Testing Service  
Princeton, NJ 08541

Dr. Ivo Molenaar  
Faculteit Sociale Wetenschappen  
Rijksuniversiteit Groningen  
Grote Kruisstraat 2/1  
9712 TS Groningen  
The NETHERLANDS

Dr. E. Muraki  
Educational Testing Service  
Rosedale Road  
Princeton, NJ 08541

Dr. Raina Nandakumar  
Educational Studies  
Willard Hall, Room 213E  
University of Delaware  
Newark, DE 19716

Academic Prog. & Research Branch  
Naval Technical Training Command  
Code N-62  
NAS Memphis (75)  
Millington, TN 38054

Dr. W. Alan Nicewander  
University of Oklahoma  
Department of Psychology  
Norman, OK 73071

Head, Personnel Systems Department  
NPRDC (Code 12)  
San Diego, CA 92152-6800

Director  
Training Systems Department  
NPRDC (Code 14)  
San Diego, CA 92152-6800

Library, NPRDC  
Code 041  
San Diego, CA 92152-6800

Liberian  
Naval Center for Applied Research  
in Artificial Intelligence  
Naval Research Laboratory  
Code 5510  
Washington, DC 20375-5000

Office of Naval Research,  
Code 1142CS  
800 N Quincy Street  
Arlington, VA 22217-5000  
(A Copies)

Special Assistant for Research  
Management  
Chief of Naval Personnel (PERS-OJUT)  
Department of the Navy  
Washington, DC 20350-2000

Dr. Judith Orasanu  
Mail Stop 239-1  
NASA Ames Research Center  
Moffett Field, CA 94035

Dr. Peter J. Pashley  
Educational Testing Service  
Rosedale Road  
Princeton, NJ 08541

Wayne M. Patience  
American Council on Education  
GED Testing Service, Suite 20  
One Dupont Circle, NW  
Washington, DC 20036

Dept. of Administrative Sciences  
Code 54  
Naval Postgraduate School  
Monterey, CA 93943-5026

Dr. Peter Pirotti  
School of Education  
University of California  
Berkeley, CA 94720

Dr. Mark D. Reckase  
ACT  
P.O. Box 168  
Iowa City, IA 52243

Mr. Steve Reise  
Department of Psychology  
University of California  
Riverside, CA 92521

Mr. Louis Rousso  
University of Illinois  
Department of Statistics  
101 Illini Hall  
725 South Wright St.  
Champaign, IL 61820

Dr. Donald Rubin  
Statistics Department  
Science Center, Room 608  
1 Oxford Street  
Harvard University  
Cambridge, MA 02138

Dr. Fumiko Samejima  
Department of Psychology  
University of Tennessee  
310B Austin Peay Bldg.  
Knoxville, TN 37966-0900

Dr. Mary Schatz  
4100 Parkside  
Carlsbad, CA 92008

Mr. Robert Semmes  
N218 Elliott Hall  
Department of Psychology  
University of Minnesota  
Minneapolis, MN 55455-0344

Dr. Valerie L. Shalin  
Department of Industrial  
Engineering  
State University of New York  
342 Lawrence D. Bell Hall  
Buffalo, NY 14260

Mr. Richard J. Shavelson  
Graduate School of Education  
University of California  
Santa Barbara, CA 93106

Ms. Kathleen Sheehan  
Educational Testing Service  
Princeton, NJ 08541

Dr. Kazuo Shigematsu  
7-9-24 Kugenuma-Kaigan  
Fujisawa 251  
JAPAN

Dr. Randall Shumaker  
Naval Research Laboratory  
Code 5500  
4555 Overlook Avenue, S.W.  
Washington, DC 20375-5000

Dr. Judy Spray  
ACT  
P.O. Box 168  
Iowa City, IA 52243

Dr. Martha Stocking  
Educational Testing Service  
Princeton, NJ 08541

Dr. William Stout  
University of Illinois  
Department of Statistics  
101 Illini Hall  
725 South Wright St.  
Champaign, IL 61820

Dr. Kikumi Tatsuoka  
Educational Testing Service  
Mail Stop 03-T  
Princeton, NJ 08541

Dr. David Thissen  
Psychometric Laboratory  
CB# 3270, Davis Hall  
University of North Carolina  
Chapel Hill, NC 27599-3270

Mr. Thomas J. Thomas  
Federal Express Corporation  
Human Resource Development  
3035 Director Row, Suite 501  
Memphis, TN 38131

Mr. Gary Thomasson  
University of Illinois  
Educational Psychology  
Champaign, IL 61820

Dr. Howard Wainer  
Educational Testing Service  
Princeton, NJ 08541

Elizabeth Wald  
Office of Naval Technology  
Code 227  
800 North Quincy Street  
Arlington, VA 22217-5000

Dr. Michael T. Waller  
University of  
Wisconsin-Milwaukee  
Educational Psychology Dept.  
Box 413  
Milwaukee, WI 53201

Dr. Ming-Mei Wang  
Educational Testing Service  
Mail Stop 03-T  
Princeton, NJ 08541

Dr. Thomas A. Warm  
FAA Academy  
P.O. Box 25082  
Oklahoma City, OK 73125

Dr. David J. Weise  
N660 Elliott Hall  
University of Minnesota  
75 E. River Road  
Minneapolis, MN 55455-0141

Dr. Douglas Wetzel  
Code 15  
Navy Personnel R&D Center  
San Diego, CA 92152-6800

German Military  
Representative  
Personalausbildung  
Koellner Str. 262  
D-5000 Koeln 90  
WEST GERMANY

Dr. David Wiley  
School of Education  
and Social Policy  
Northwestern University  
Evanston, IL 60201

Dr. Bruce Williams  
Department of Educational  
Psychology  
University of Illinois  
Urbana, IL 61801

Dr. Mark Wilson  
School of Education  
University of California  
Berkeley, CA 94720

Dr. Eugene Winograd  
Department of Psychology  
Emory University  
Atlanta, GA 30322

Dr. Marvin F. Wittoff  
PI RSI-REC  
88 Pacific St., Suite 4556  
Monterey, CA 93940

Mr. John H. Wolfe  
Navy Personnel R&D Center  
San Diego, CA 92132-6800

Dr. Kentaro Yamamoto  
63-07  
Educational Testing Service  
Rosedale Road  
Princeton, NJ 08541

Ms. Duanli Yan  
Educational Testing Service  
Princeton, NJ 08541

Dr. Wendy Yen  
CTB/McGraw-Hill  
Del Monte Research Park  
Monterey, CA 93940

Dr. Joseph L. Young  
National Science Foundation  
Room 320  
2001 G Street, N.W.  
Washington, DC 20550